DOCUMENT RESUME

ED 400 268 TM 025 373

AUTHOR Bode, Rita K.

TITLE Is It Ability Grouping or the Tailoring of Instruction That Makes a Difference in Student Achievement?

PUB DATE Apr 96

NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Ability Grouping; Elementary Education; *Elementary School Students; *Equal Education; Grade 8; High Achievement; Item Response Theory; Low Achievement; *Mathematics Achievement; Mathematics Instruction; *Track System (Education)

IDENTIFIERS Hierarchical Linear Modeling; Rasch Model; Second International Mathematics Study

ABSTRACT

The controversy about ability grouping in education boils down to a conflict between the educational goals of excellence and of equity. There is considerable evidence that ability grouping is effective in producing learning, but not for all students. This study addresses substantive issues raised in previous research, exploring whether ability grouping works, and for whom it works. The focus of the study is the use of one type of ability grouping, within class ability grouping, in eighth-grade mathematics instruction. Data from the Second International Mathematics Study were used, with a sample of 3,991 U.S. eighth graders from 127 schools. Student- and class-level data files were created, and Rasch measures were developed. The statistical model used was hierarchical linear modeling. Results show that the use of within-class ability grouping and instructional tailoring has no effect on average eighth-grade mathematics achievement. In addition, the use of within-class grouping has no effect on the link between previous and subsequent eighth-grade mathematics achievement and the extent to which it varies across classes. It appears that instructional tailoring allows high achievers to perform to their maximum but does not have negative effects on low achievers. (Contains 2 tables, 6 figures, and 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 268

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
RITA K. BODE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Is it Ability Grouping or the Tailoring of Instruction That Makes a Difference in Student Achievement?

Rita K. Bode

University of Illinois at Chicago

Paper presented at the Annual Meeting of the
American Educational Research Association
New York NY, 1996

TM025373

BEST COPY AVAILABLE

Background

There has been a longstanding controversy concerning the grouping of students by ability for instruction. Arguments favoring its use have stressed benefits for both teachers and students with its use. Teachers are thought to benefit because too much heterogeneity may create instructional problems and it is felt that there are limits on the amount of class heterogeneity that a teacher can reasonably handle (Good and Marshall, 1984). Students, regardless of their ability level, are thought to learn more when the pace and content of instruction are matched to their level (Slavin, 1987). When taught in ability groups, high achievers move at a faster pace, are exposed to more material, and, therefore, learn more than if tied to the pace of the class as a whole. Low achievers, on the other hand, move at a slower pace but, even though they are exposed to less material, learn more of what they are taught than if tied to the pace of the class as a whole (Sorensen and Hallinan, 1986).

Arguments opposing its use have stressed the inequitable effects of ability grouping for students. Its effects are thought to be beneficial for high achievers but detrimental for low achievers; that is, high achievers learn more than low achievers in a given amount of time. Among others, Oakes (1985) and Page (1992) claim that students in low ability groups are exposed to less of the curriculum and that presentations are watered down and consist of lower level skills. Ability grouping is also thought to increase social stratification and lead to self-fulfilling prophecies regarding teachers expectations of lower ability students (Rosenbaum, 1980; Eder, 1981).

Some of these contradictory conclusions can be attributed to the fact that different outcomes--academic versus social--are being referenced. However, even when academic outcomes are the focus, contradictory results are found. One reason may be that different issues are being addressed by proponents and opponents. Proponents typically compare the performance of students who are ability grouped for instruction to that of students whose instruction is provided to the class as a whole (i.e., Hallinan and Sorensen, 1987). This research tends to be quantitative in nature and the outcome is typically some measure of student achievement. Its purpose tends to be a determination of the effectiveness of ability grouping in producing student achievement as compared to not using it--usually without regard to differential effects across ability groups. Opponents typically compare the performance of students in the higher ability groups to that of students in the lower ability groups (i.e., Oakes, 1985). This research tends to be qualitative in nature and it tends to describe the quality of instruction across ability groups--usually without regard to its effect on student achievement. As a consequence, conclusions drawn from these two types of research have been contradictory in terms of whether or not ability grouping can be considered a good practice or a bad one.

The controversy boils down to a conflict between two goals of education: excellence (or productivity) and equity (or inequality) (Oakes, Gamoran, and Page, 1992). Working toward the goal of excellence, practices that result in the highest achievement for students overall would be desirable; working toward the goal of equity, practices that reduce the gap in achievement between high achievers and low achievers would be desirable. In order to satisfy both goals, ability grouping would need to be as, if not more, effective for all students; that is, achievement would need to be as high as or higher for both high and low achievers in classes in which ability grouping is used.

There is considerable evidence that ability grouping is effective in producing learning, but not for all students (Good and Marshall, 1984). When its effects differ, it may be that some types of ability grouping are effective while others are not, or ability grouping may be more effective in certain grades and subject areas, or ability grouping may be differentially effective for different types of students. In the first two instances, the issue is one of excellence and knowing what type of ability grouping is best to use, when, and for which subjects is important. In the third instance, the issue is one of equity and understanding why and how these differential effects occur is important.

Purpose of Study

This study addresses substantive issues raised in previous ability grouping and other school effects research. It not only determines if ability grouping works but also for whom it works; that is, does the use of ability grouping produce higher achievement for students overall and do both high and low achievers benefit from its use. In terms of excellence, this study determines if the use of ability grouping has an effect on average student achievement, regardless of ability level or other student characteristics. In terms of equity, this study determines if the use of ability grouping has a differential effect on students of different ability levels.

The focus of this study is the use of one type of ability grouping, within-class ability grouping, in eighth-grade mathematics instruction. Typically at the eighth-grade level, if ability grouping is used, it is between-class ability grouping or tracking; thus these are the types of ability grouping typically studied at this grade level. Despite the common use of between-class ability grouping or tracking, teachers at the eighth grade are still likely to create instructional groups to further narrow the range of ability within classes, especially classes in which the subject matter is hierarchical, as it is in mathematics. Thus, the effects of within-class ability grouping need to be examined.

When within-class ability grouping is used in a classroom, teachers may differ in the extent to which they tailor instruction to student needs. For this reason, in this study ability grouping is measured not only by the extent to which the grouping arrangement is used in classrooms but also by the extent to which the curriculum is differentiated across instructional groups. Hidden in the responses to the direct question of whether ability grouping was used in classroom instruction were many variations in practice. Some teachers may have pulled out the most able students and provided different instruction to them than was provided to the other students while others may have pulled out the least able students and instructed them separately. Some teachers may have formed instructional groups that differed in terms of ability occasionally or for some topics while others may have used the same organizational grouping of students for all instruction. Yet they all might have reported using ability grouping in their instruction.

It isn't enough to ask whether ability grouping is used in a classroom. According to Gamoran (1987), grouping does not produce achievement; instruction does. Thus one needs to examine the consequences of the grouping arrangement used in instruction. Once the teacher forms instructional groups, how is instruction provided? Do teachers vary the content of instruction, the pacing of instruction, or the assignments that are given to reinforce instruction? Do teachers tailor instruction to student needs frequently, occasionally, rarely, or never?

Taking advantage of advances in statistical techniques that model multilevel data, this study does not directly measure the effects of grouping on achievement status but instead determines its effects on growth in mathematics achievement. It uses a theory-based model of learning (Sorensen and Hallinan, 1977) in which student characteristics known to influence student achievement are included. It models data collected at two organizational levels--student and class--by determining: 1) the effects of student characteristics on mathematics achievement, 2) whether average student achievement and the effects of student characteristics vary across classes, and 3) whether ability grouping and instructional tailoring, controlling for other classroom characteristics, explain the differences in average mathematics achievement across classes. This study incorporates the use of Rasch measurement, a more accurate measurement of the variables under investigation. As estimates of previous and subsequent student achievement, Rasch measures are capable of estimating underlying growth in mathematics and provide more accurate estimates of the other student and class characteristics than traditional indicators do. Finally, the use of Rasch measures for ability grouping and instructional tailoring avoids problems resulting from different understandings of the meaning of these terms encountered when using dummy variables, more accurately describes each practice, and provides more information about their use.

METHODOLOGY

Extant data were used in this study to test the statistical models. Benefits of using such a dataset include the availability of a large and representative sample and a broad range of data that could be used to create the variables of interest. Care was taken to select the dataset that best provided the raw materials needed to developed measures of these variables.

Sample

To develop measures of ability grouping and instructional tailoring that describe practice within a classroom, information on classroom processes are needed. Also while the analytic technique used in this study could have been run on a smaller sample, more accurate modeling required a sample in which there were a substantial number of units (classes) and cases (students) within each of these units. One of the few national databases that met these requirements was the Second International Mathematics Study (SIMS). In SIMS, schools were sampled such that they were representative of schools nationally in terms of geographic area and school size. Depending on school size, one or two intact classrooms were randomly sampled from each school. As a result, students within intact classrooms, and their teachers, were included in the study (Westbury, Caroll, & Thalathoti, 1989).

SIMS was undertaken by the International Association for the Evaluation of Educational Achievement to test mathematics achievement of national samples of students in 20 countries on common tests. The testing was conducted between 1980 and 1982. SIMS contained two levels of population: one consisted typically of students in the national class in which the model age was 13 and the other consisted of students in the terminal secondary grades of each system. The first level, eighth-graders in the U.S. sample, is used in this study.

The SIMS study sampled 6935 U.S. eighth-grade students from 164 schools who were in 235 classes whose main mathematics subject matter was characterized by the teacher as remedial, typical, enriched, or algebra. The sample used in this study was composed of students within these classes for which complete data for both the student- and class-level variables included in the model were available. This selection criteria resulted in a sample of 3991 students from 215 remedial, typical, enriched, and algebra classes in 127 schools.

Instrumentation

Items of interest were selected from the SIMS database and used to create student- and class-level data files. To develop Rasch measures, the items were analyzed using the Rasch software, BIGSTEPS, version 2.56. Estimates of the reliability of each of these scales, where appropriate, are incorporated in the description of the operationalization of the variables in the model and are presented below.

The indicators for the student-level variables are as follows:

Gender	Dummy variable: 0 = female; 1 = male
Minority Group Status	Dummy variable: 0 = minority; 1 = majority
Socioeconomic Status	Rasch measure composed of mother's and father's education and occupation (person separation reliability = .73)
Motivation:	Rasch measure of responses to ten attitudinal items (person separation reliability = .81)
Opportunity-to-learn:	Rasch measure composed of the item content to which students indicated they were exposed (person separation reliability = .89)
Previous Achievement:	Rasch measure based on responses to items in the fall math test (person separation reliability = .87)

The indicators of the class-level variables are as follows:

Class Gender Mix:	Proportion of males within the class
Class Minority Group Mix:	Proportion of minority students within the class
Class SES Level:	Average student SES Rasch measure
Class Motivation Level:	Average student motivation Rasch measure
Class Opportunity-to-learn Level:	Average student opportunity-to-learn Rasch measure
Class Ability Level:	Average fall math test Rasch measure
Class Ability Range:	Standard deviation of the fall math test Rasch measure
Grouping Level:	Rasch measure of the extent to which grouping is used (person separation reliability = .83)
Tailoring Level:	Rasch measure of the extent to which tailoring is used (person separation reliability = .72)

Measuring Grouping and Tailoring

Of all the measures that need to be created to study the impact of ability grouping on student achievement, the measures of ability grouping and instructional tailoring use in the classroom are the most crucial. Because ability grouping and tailoring could have meant different things to different people, identifying the extent to which characteristics of these practices were present in a classroom, rather than just whether or not teachers reported using them, was important.

What is needed to define a continuum are descriptors of behaviors or attributes that represent more or less of a behavior or attribute. When one creates a scale that describes the extent to which ability grouping is used in a classroom, it is possible to describe actual practice within the classroom. Such a scale depends on identifying aspects of classroom practice that characterize ability grouping. More precision can be achieved by using a scale wherein items are positions along a continuum that define the construct. Regardless of how the item response scales are defined (e.g., frequency of occurrence, degree or amount of satisfaction or agreement, etc.), these items need to form a continuum in which one end represents characteristics that are rare within classrooms and the other end represents characteristics that are common within classrooms.

Once this item continuum has been created, it is possible to position teachers along the same continuum such that teachers at one end respond negatively to items which characterize the use of ability grouping (and therefore exhibit few characteristics of ability grouping in their classrooms) and those at the other end respond positively to such items (and therefore exhibit many characteristics of ability grouping in their classrooms).

To create a measure of the extent of use of ability grouping in mathematics classrooms, two types of items concerning classroom grouping arrangements were used: one dealing with the amount of class time spent in different classroom arrangements and the time students spent in various activities and the other dealing with situations occurring regularly in within classrooms in which small group instruction was used. The first set of items consisted of time spent in: whole class lecture or discussion, small group instruction, all students working individually, doing seatwork or blackboard work, listening to lectures or explanations, and working in small groups. The second set of items identified the type of grouping arrangement used in the classroom: most able students working separately while the rest of the class worked as a single group, least able students working separately while the rest of the class worked as a single group, class split into three or more groups each at a different ability level, none of the above occurring regularly, or question does not apply--small group instruction not used. Despite the use of different response scales in these items, a single measure of

the extent to which ability grouping was used could be created; the continuum ranged from practices which were rarely used in eighth-grade mathematics classroom to practices which were commonly used in these classrooms. The grouping continuum is illustrated in the first column in Figure 1.

To create a measure of the extent of use of instructional tailoring in mathematics classrooms, two types of item concerning tailoring were used: one dealing with beliefs about the importance of tailoring and the other dealing with actual tailoring practice. The first set of items consisted of beliefs concerning the importance of: giving less able students assignments that are simple enough that they can progress without making mistakes, assigning problems which require the abler students to do more than follow examples that have already been demonstrated, varying the difficulty of questions posed in classroom discussion, giving abler students assignments with some problems which are truly difficult for them to solve, and giving assignments which are tailored to the particular instructional needs of individual students. The second set of items concerned actual classroom practice: teachers were asked how often some of their students were asked to do exercises or problem assignments which were different from those given other students in the class and other characteristics of their teaching (teaching all students the same content but letting them proceed at their own pace, varying the content across students or groups of students, assigning all students the same set of exercises but varying the date of completion from student to student, assigning exercises to some students that other students in the class would not be expected to do) and how assignments differ across students (assigning more exercises to some students than other students, assigning more difficult exercises to some students than other students, and assigning exercises on topics to some students which other students aren't expected to cover this year). Again, although different response scales were used in these items, a single measure of the extent to which instructional tailoring was used within a classroom could be created; the continuum ranged from beliefs or practices that were relatively rare in eighth-grade mathematics classrooms to those that were more common in these classrooms. The tailoring continuum is illustrated in the second column in Figure 1.

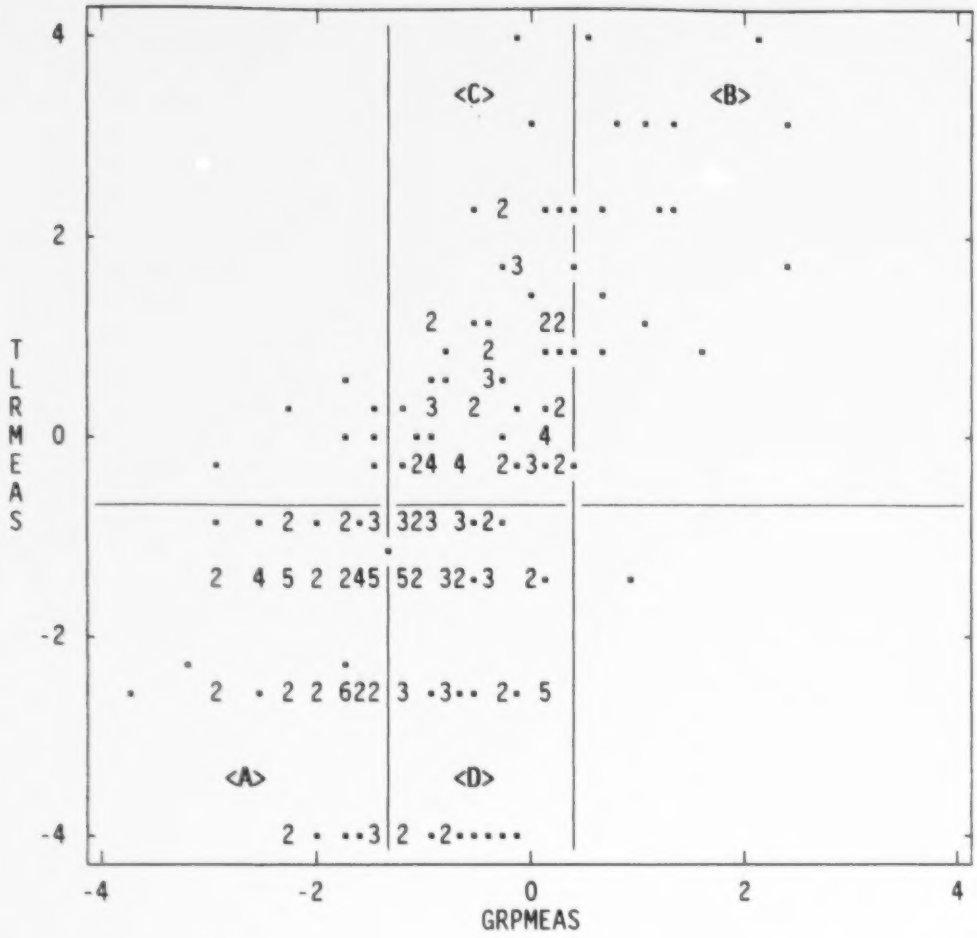
Relationship Between Grouping and Tailoring Measures

Are grouping and tailoring practice simply different indicators of the same construct or are they two different constructs? To answer this question, two Rasch calibrations were run. In one calibration, both types of items were combined to create a single measure, and in separate calibrations, the items were split to create two measures. The results of these two sets of calibrations were compared to determine which provided the most useful description of these class variables.

A comparison of the results of the separate and combined calibrations shows person separation is greater when the two sets of items are combined but item fit is better when the two sets of items are calibrated separately. Because there was no consensus using these criteria in terms of which calibration is better, a plot of teachers' positions on both measures was used to illustrate the relationship between these two measures. If the plot points fall closely along the identity line (diagonal), that would indicate that teachers' positions on one scale are the same as their position on the other scale; that is, the extent of their use of ability grouping is the same as the extent to which they use instructional tailoring.

The relationship between the grouping and tailoring can be seen by plotting teachers' measures on these two variables against each other. Because the measures are based on a common calibration of both types of items, teachers' positions on these continua can be directly compared. Further, using Rasch calibrations, teacher measures and item calibrations are positioned along the same continuum, so it is possible to use the position of items on the continuum to describe teachers' practice and see how they are distributed in terms of grouping and tailoring. The plot of teachers' grouping and tailoring measures is presented in Figure 2. As illustrated by the plot, the relationship between the extent to

Figure 1. Item Map for Grouping and Tailoring Measures



Legend:

<A> = neither ability grouping nor tailoring used
<B> = both ability grouping and tailoring used
<C> = tailoring and some form of grouping used
<D> = no tailoring but some form of grouping used

Figure 2. Plot of Teacher Grouping and Tailoring Measures

which grouping and tailoring are used was only moderate; in fact, the Pearson's product moment coefficient was only .551. This level of relationship indicates that these two practices--the extent to which teachers use ability grouping and the extent to which they tailor instruction--were neither completely separate constructs nor were they essentially the same construct. If they had been completely separate, two measures would clearly have been called for; if they had been essentially the same, one measure would have sufficed.

The practice of teachers in this study can be described initially in terms of the types of grouping arrangements used and the frequency of their use of tailoring. Lines are drawn to distinguish between different groups of teachers. These lines represent the gaps in the item calibrations as shown in Figure 1. There is a gap on the grouping continuum between 0 and 1; this gap distinguishes teachers who group their most able students together from those who use traditional ability grouping or group their least able students together. Another gap along the grouping continuum is between -1 and -2; this gap distinguishes between teachers who group their most able students together from those who don't group at all or use mixed ability grouping. Along the tailoring continuum there is only one gap, falling between 0 and -2; this gap distinguishes teachers who rarely or never tailor from those who use it at least occasionally.

Once the lines representing gaps in the continua that distinguish teachers by their grouping and tailoring practice are added to this plot, it is possible to identify main four groups of teachers in terms of their practice. The first group, indicated on the plot by an <A>, are teachers who use neither ability grouping nor tailoring. The second group, indicated on the plot by a <B>, are teachers who use both ability group and tailoring. These two groups of teachers are at different ends of the grouping and tailoring continua and represent the expectation in terms of the assumed close connection between ability grouping and instructional tailoring--that expectation being that teachers who group by ability also tailor instruction to the needs of the students within the groups.

However, there are other teachers who don't fit this pattern. Teachers who group their most able students together have tailoring practices that span the entire range of the scale. The third group, indicated on the plot by a <C>, are teachers who group their most able students together and tailor their instruction. The fourth group, indicated on the plot by a <D>, are teachers who group their most able students together but do not tailor their instruction. Finally, there are a few teachers who tailor their instruction but do not use ability grouping and there is one teacher who uses ability grouping but not tailoring. These anomalies will not be dealt with in this discussion.

What are the characteristics of teachers who fall into these four groups? Sixty-three teachers (29.3%) fall into group <A>; these teachers appear to concentrate on whole class instruction or other arrangements that do not entail grouping; their students spend less than 10% of their time in group work; they believe tailoring is of some importance but rarely or never tailor their instruction. In contrast 18 teachers (8.4%) fall into group <B>; these teachers appear to use traditional ability grouping or group their least able students together, spend more than 50% time in small group instruction, believe tailoring is of utmost importance, tailor their instruction frequently, and tailor both their assignments and their instruction. Other characteristics of teachers were examined to investigate if they played a part in determining a teacher's grouping and tailoring practice. Teacher gender, age, education, and experience were examined and no significant differences were found across groups.

What about the teachers who use grouping to the same extent but differ in their use of tailoring? Sixty-six teachers (30.7%) fall into group <C>; they tend to group their most able students together, spend less than 50% time in small group instruction, believe tailoring is of utmost importance and use it frequently. These teachers are similar to their colleagues in group <B> but use a less extreme form of grouping; instead of grouping all of their students by ability, they separate only their most able students and have them work together. Finally, 60 teachers (27.9%) fall into group <D>; these

teachers are like those in group <C> in the type of ability grouping used but they believe tailoring is only of some importance and rarely or never use it. The lack of a relationship between grouping and tailoring for these two groups of teacher explains the low correlation between use of grouping and tailoring found in this study.

Analysis

The statistical model used in this study was Hierarchical Linear Modeling (HLM), an analytic technique developed by Raudenbush and Bryk (1986) in which two levels of data--student and class--are used to model student achievement. The program used to analyze these multilevel data was HLM2 (Version 3.0), developed by Bryk, Raudenbush, and Congdon (1993). Three HLM models were estimated and tested. Between- and within-class variability are decomposed in the initial model and subsequent models were used to explain this between- and within-class variance. The second model included only student characteristics and the third model included class characteristics: grouping and tailoring and the contextual variables.

Specifically, the first (baseline) HLM model determined the total amount of variance in mathematics achievement and the proportion that was accounted for between and within classes. The second (within-class) HLM model introduced the following variables to the model to determine if these student characteristics had a significant effect on student achievement: gender, minority group status, family socioeconomic status, motivation, opportunity-to-learn, and previous achievement. The significance level used in the within-class HLM model was .01 because of the large sample size (3991). In the within-class model, HLM also determined whether the effects of student-level variables were fixed or random, that is, whether the effects varied across classes. Only those variables whose effects varied across classes were included in subsequent models. Thus, class-level variables were used to model not only on average student achievement but also the relationship between previous and subsequent achievement. The third (between-class) HLM model added the grouping and tailoring variables and the contextual variables which were aggregates of student-level variables to determine the effect of grouping and tailoring once the composition of the classroom was taken into account. Class ability level and range, socioeconomic level, opportunity-to-learn level, motivation level, and the proportion of males and minority group students in the class were added to the model at this stage. The significance level used in the between-class HLM models was .05 because of the smaller sample size (215).

For ease of interpretation, except for the dummy variables, student- and class-level variables were centered around their means using a rationale developed by Gamoran (1991). For student-level variables that were fixed, that is, they did not vary across classes, the values were centered around the grand mean; for student-level variables that were random, that is, they did vary across classes, the values were centered around their class mean. For class-level variables, the values were centered around the grand mean. Using this scheme, the intercept can be interpreted as the average student achievement for students who have average levels of the student characteristics, or are coded "0" in the dummy variables. To summarize the results of this series of models, the percentage of variance accounted for by each model was determined. At each level it was possible to quantify the contribution of variables to the modeling of mathematics learning.

RESULTS

Three HLM models were used to answer the research questions dealing with the effects of ability grouping and instructional tailoring on student achievement. In the baseline model (not shown), variance in student achievement is decomposed into between- and within-class variability. The results indicate that 54% (15.171) of the variability in posttest mathematics performance is due to variability

across classes with 46% (12.938) due to variability within classes. Both variance estimates are the baseline for subsequent proportions of variance explained.

Estimation of the student-level model is presented in Table 1 under the heading "Within-Class Model." In this model, student-level variables are used to predict individual achievement. The values presented for this model represent the average of the coefficients across classes. The student-level variables that significantly predict posttest mathematics performance are as follows: minority group status (MAJMIN), socioeconomic status (SESMEAS), motivation (MTVMEAS), opportunity-to-learn (OTLMEAS), and previous math achievement (PREMEAS). The only student-level variable included in the model that does not influence math achievement is gender (GENDER). In other words, majority group students with high levels of previous achievement, motivation, opportunity to learn, and socioeconomic status are the students who have high posttest mathematics performance. The results for the test for variability in the intercepts across classes for the student-level model (not shown) indicate that adding the student-level variables to the model reduces the variability of within-class student achievement 48.7%.

The results of the between-class model are presented in Table II. With the inclusion of these contextual variables, the effects of grouping and tailoring on average student achievement are not significant. This indicates that it is not the practice of grouping or tailoring per se that influences average class achievement but the class context. Class-level variables that have a significant effect on variability in means across classes are: class ability (AVGPRE) and opportunity-to-learn (AVGOTL) levels indicating that classes in which posttest mathematics performance is highest are those in which class ability and opportunity to learn are highest. The class-level variables that have a significant effect on the relationship between previous and subsequent achievement are: the extent to which tailoring is used in the class (TLRMEAS), class ability (AVGPRE), and class opportunity-to-learn (AVGOTL) indicating that the link between previous and subsequent achievement is highest in classes in which tailoring is used extensively and class ability and opportunity to learn are high.

Table I

HLM Results for Student-Level Model

	Within-Class Model
INTERCEPT	20.106***
GENDER	0.008
MAJMIN	0.536***
SESMEAS	0.079**
MTVMEAS	0.474***
OTLMEAS	0.128**
PREMEAS	2.838***

Legend: * = .05; ** = .01; *** = .001

Table II

HLM Results for Class-Level Model

	Between-Class Model
INTERCEPT	20.379***
GRPMEAS	0.077
TLRMEAS	0.060
PROPMAL	-0.011
PROPMIN	-0.005
AVGSES	0.121
AVGMTV	0.164
AVGOTL	0.504**
AVGPRE	3.974***
SDPRE	0.367
PREMEAS	3.007***
GRPMEAS	-0.080
TLRMEAS	0.232***
PROPMAL	-0.007
PROPMIN	-0.005
AVGSES	-0.165
AVGMTV	-0.179
AVGOTL	0.179*
AVGPRE	0.464**
SDPRE	-0.063

Legend: * = .05; ** = .01; *** = .001

The residual parameter variances when grouping, tailoring, and class contextual variables (not shown) indicate that adding these variables reduces the variability across classes as follows: intercept, 85.1%; and previous achievement, 14.2%. Even after these variables have been added to the model, average achievement still varies across classes and there is still variation in the effect previous on subsequent achievement. The large reduction in variance for the intercept indicates that these contextual variables, predominantly average ability and opportunity to learn, explain by far most of the variability in average achievement across classes whereas they explain little of the variability across classes in the relationship of subsequent achievement with previous achievement.

Discussion

The results show that grouping has no effect on differences in average mathematics achievement across classes. One way to illustrate the effects of ability grouping on mathematics achievement is to calculate the average scores of students in different types of grouping arrangements using the coefficients obtained from the between-class model and the Rasch measures for the grouping variable (see Figure 1). Because the Rasch calibrations for items and persons are on the same continuum, the

location of items on this scale can be used to describe teachers on this same scale. In this case, teachers who had grouping measures around + 1 are identified as teachers who used traditional ability grouping in which students were assigned to one of three or more groups within which ability level differed or who separated their less able students to work together; those who had grouping measures around 0 are identified as teachers who separated their more able students to work together; and those who had grouping measures around -2 are those who did not use grouping at all or used mixed ability grouping.

When class context is taken into account and the coefficients from the between-class model are used in the calculations, there is no difference in average achievement. All things being equal, that is, controlling for all the other class characteristics included in the model, the average achievement of students in classes in which traditional ability grouping is used or teachers group their least able students together is 20.456 [20.379 + (1).077]; for students in classes in which the most able students work together, the average is 20.379 [20.379 + (0).077]; and for students in classes in which grouping is not used at all or mixed ability grouping is used, average achievement is 20.225 [20.379 + (-2).077]. Figure 3 illustrates average mathematics achievement of classes using different grouping arrangements.

The results further show that tailoring also has no effect on differences in average mathematics achievement across classes. Using the Rasch measures for tailoring (see Figure 1) and the tailoring coefficients from the between-class model, it is possible to estimate the average scores of students in classes in which tailoring is used to different extents. Teachers who have tailoring measures around 2 can be described as using tailoring frequently; those who have tailoring measures around 0 can be described as using tailoring occasionally, and those who have tailoring measures around -2 can be described as never or rarely using tailoring.

When class context is taken into account and the coefficients from the between-class model are used, all things being equal, the results are as follows: in classes in which tailoring is used frequently, the average is 20.499 [20.379 + (2).060]; in classes in which tailoring is used occasionally, average achievement is 20.379 [20.379 + (0).060]; and in classes in which tailoring is used rarely or never, the average is 20.259 [20.379 + (-2).060]. Figure 4 illustrates the average mathematics achievement of classes in which tailoring is used with different frequency.

The effects of grouping and tailoring were also determined on differences in the relationship between previous and subsequent achievement. Whereas the findings concerning the effects of grouping and tailoring on average achievement address the issue of excellence, the findings concerning their effects on the relationship between previous and subsequent achievement address the equity issue. The results of this study show that grouping has no effect on the relationship between previous and subsequent achievement but tailoring does. To illustrate the effects of ability grouping and tailoring on the gap in posttest mathematics achievement between high and low achievers, the average measures of students who differed in terms of pretest achievement level and the extent to which grouping and tailoring was used in their class are obtained.

Average pretest and posttest measures were calculated for four groups of students: those in classes in which traditional ability grouping was used (grouping measure greater than one) who scored at least one standard deviation above the pretest mean (n = 51, pretest mean = 25.078, pretest standard deviation = 1.534, posttest mean = 27.373, posttest standard deviation = 3.039); those in classes in which ability grouping was used (grouping measure greater than one) who scored at least one standard deviation below the pretest mean (n = 148, pretest mean = 12.581, pretest standard deviation = 1.552, posttest mean = 15.682, posttest standard deviation = 2.900); those in classes in which ability grouping was not used (grouping measure less than -2) who scored at least one standard deviation above the pretest mean (n = 224, pretest mean = 27.036, pretest standard deviation = 2.712, posttest

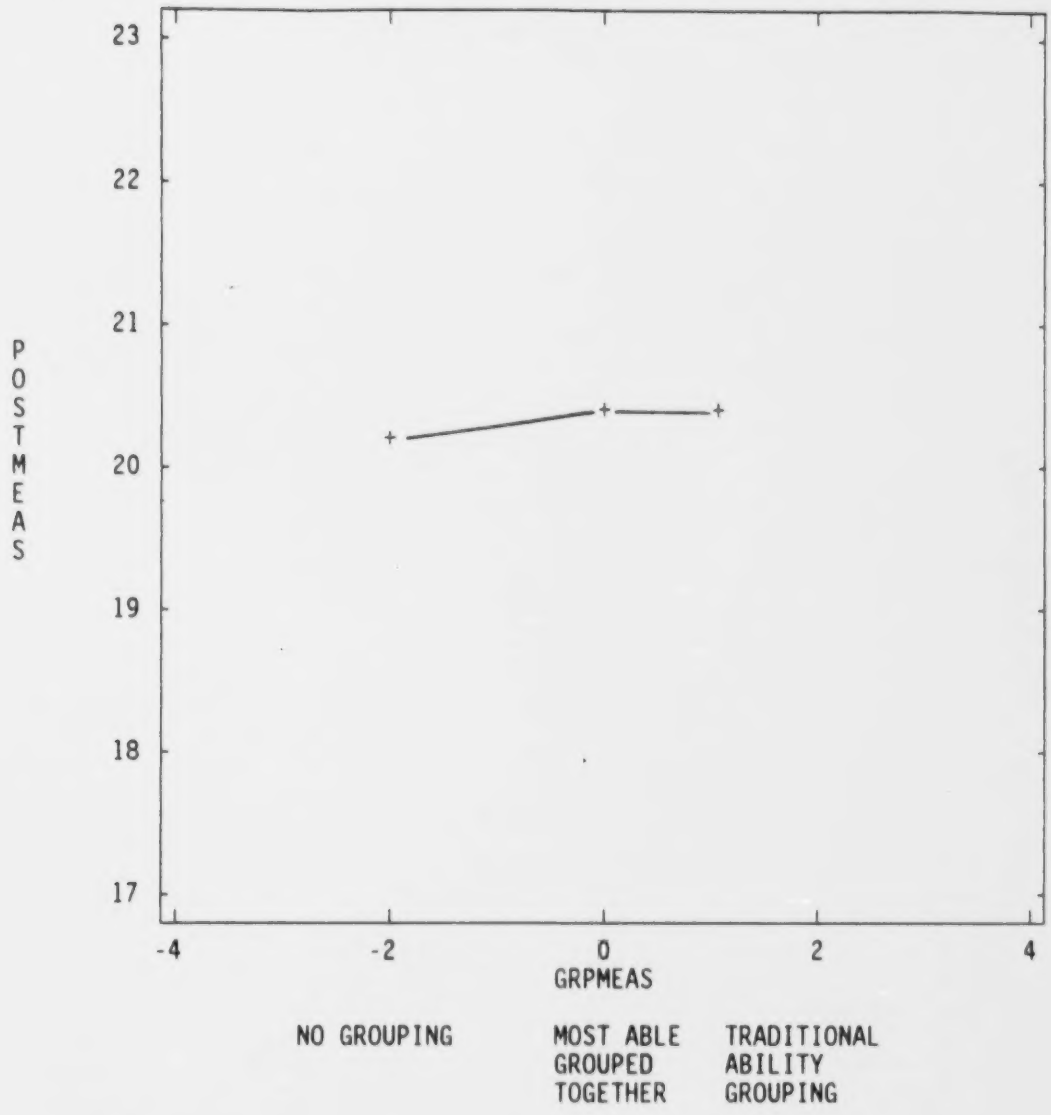


Figure 3. Posttest Measures for Different Levels of Grouping

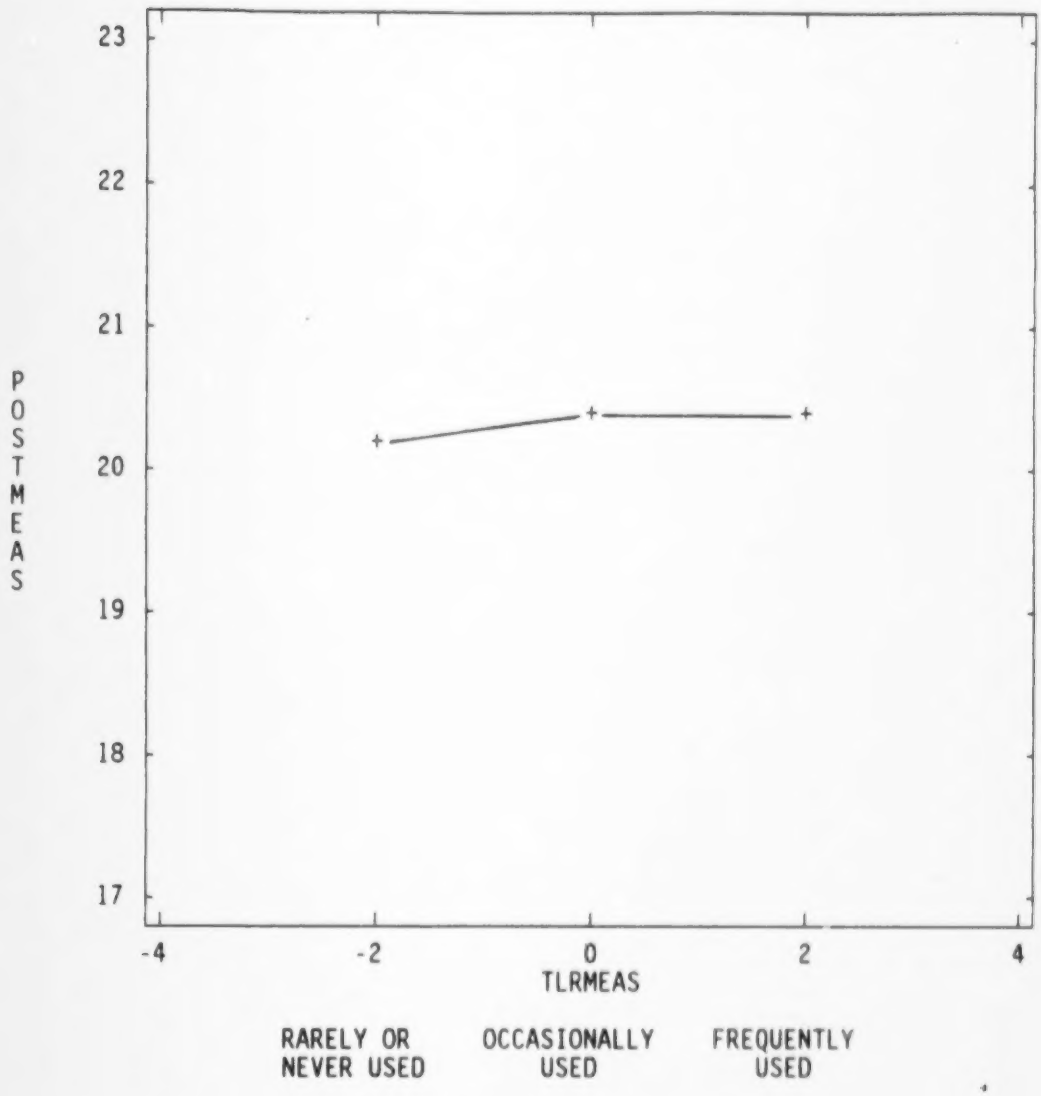


Figure 4. Posttest Measures for Different Levels of Tailoring

mean = 29.330, posttest standard deviation = 3.847); and those in classes in which ability grouping was not used (grouping measure less than -2) who scored at least one standard deviation below the pretest mean (n = 118, pretest mean = 12.585, pretest standard deviation = 1.676, posttest mean = 15.008, posttest standard deviation = 3.084).

Since the pretest means for the grouping versus nongrouping students are different, posttest averages are adjusted to compensate for this difference; that is, the adjusted posttest measures are estimates of what the posttest measures would have been if the two groups were initially comparable. For the high achievers, the pretest average for the nongrouping students was 1.958 higher than that of the grouping students. For the low achievers, the pretest average for the grouping students was .004 higher than that of the nongrouping students. A simple adjustment was made in the posttest averages to correct for these differences. The posttest average for high achieving grouping students was increased by 1.958 and the posttest average for low achieving nongrouping students was increased by .004; thus the adjusted posttest average for high achieving grouping students was estimated as 29.331 and the adjusted posttest average for low achieving nongrouping students was estimated as 15.686.

Adjusted posttest measures for the high and low achieving grouping and nongrouping students are presented in Figure 5. This figure illustrates that, when differences in pretest achievement levels are taken into account, the difference in adjusted posttest mathematics achievement of high and low achieving students is essentially the same in classes in which ability grouping is used as in classes in which it is not used.

Similarly average pretest and posttest measures were calculated for four groups of students who differed in terms of the extent to which tailoring was used in the class and the ability level of the students: students in classes in which tailoring was used frequently (tailoring measure greater than zero) who scored at least one standard deviation above the pretest mean (n = 154, pretest mean = 25.597, pretest standard deviation = 1.784, posttest mean = 27.968, posttest standard deviation = 4.091); those in classes in which tailoring was used frequently (tailoring measure greater than zero) who scored at least one standard deviation below the pretest mean (n = 147, pretest mean = 12.673, pretest standard deviation = 1.575, posttest mean = 15.878, posttest standard deviation = 3.302); those in classes in which tailoring was rarely or never used (tailoring measure less than -2) who scored at least one standard deviation above the pretest mean (n = 122, pretest mean = 26.672, pretest standard deviation = 2.291, posttest mean = 28.631, posttest standard deviation = 3.507); and those in classes in which tailoring was rarely or never used (tailoring measure less than -2) who scored at least one standard deviation below the pretest mean (n = 178, pretest mean = 12.522, pretest standard deviation = 1.698, posttest mean = 15.478, posttest standard deviation = 2.596).

Since the pretest means for the tailoring versus nontailoring students also were different, posttest averages were adjusted to compensate for these differences. For the high achievers, the pretest average for the nontailoring students was 1.075 higher than that of the tailoring students. For the low achievers, the pretest average for the tailoring students was .151 higher than that of the nontailoring students. The posttest average for high achieving tailoring students was increased by 1.075 and the posttest average for low achieving nontailoring students was increased by .151; thus the adjusted posttest average for high achieving tailoring students was estimated as 29.043 and the adjusted posttest average for low achieving nontailoring students was estimated as 15.629.

Average adjusted posttest measures for the high and low achieving tailoring and nontailoring students are presented in Figure 6. When differences in pretest achievement levels are taken into account, the figure illustrates that the difference in adjusted posttest mathematics achievement of high and low achieving students is greater in classes in which tailoring is frequently used than in classes in which

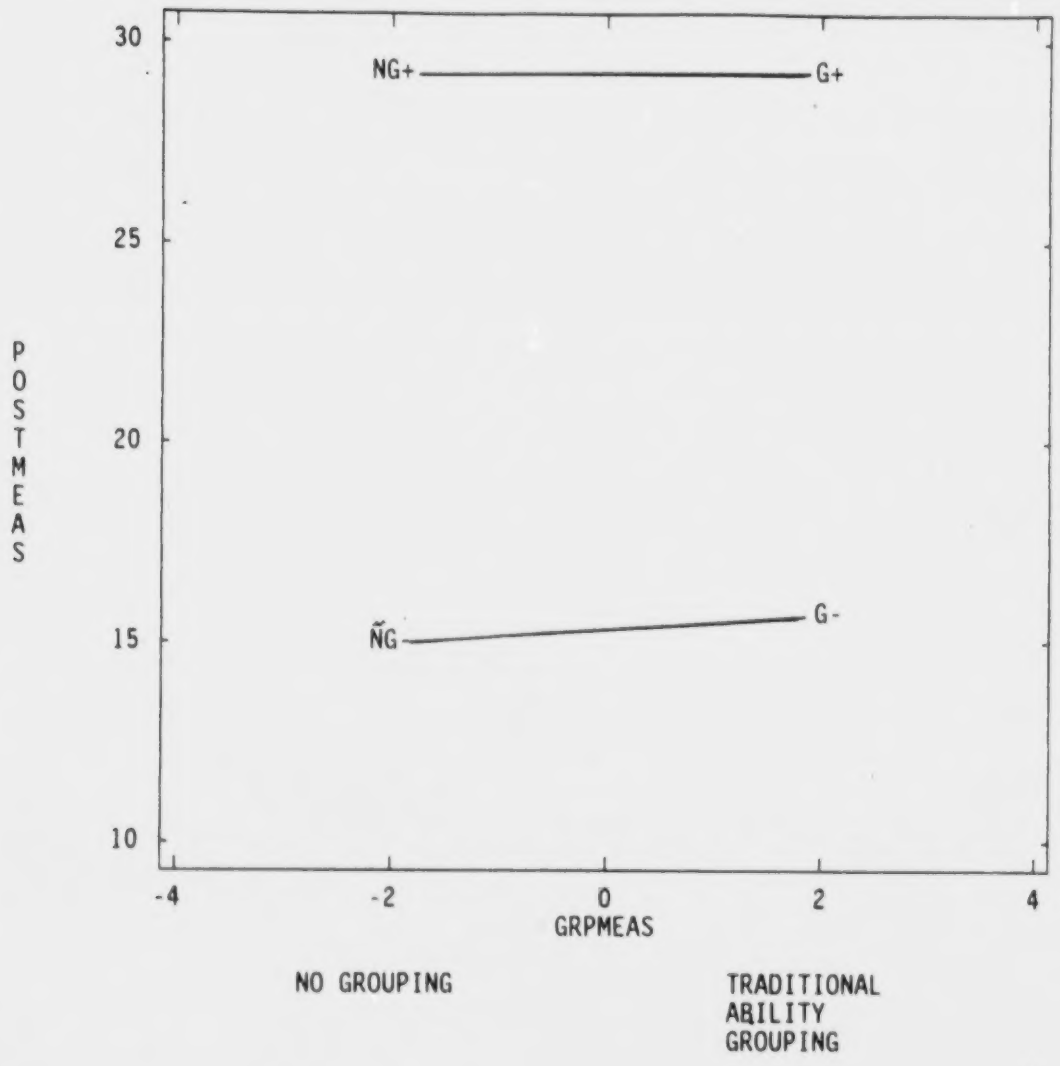


Figure 5. Posttest and Adjusted Posttest Measures for Different Levels of Grouping for High and Low Achievers

tailoring is rarely or never used. Despite this greater gap, neither high achievers nor low achievers appear to suffer from more extensive use of tailoring: high achievers benefit from its use while low achievers perform essentially the same whether or not tailoring is used.

Conclusions

The focus of this study was not only if but for whom ability grouping had an effect and what that effect was. The effects of within-class ability grouping and instructional tailoring on eighth-grade mathematics achievement were considered in terms of two goals of schooling: excellence and equity. For ability grouping and tailoring are to be considered effective and equitable, both goals would need to be met.

From these results one can conclude that, once characteristics of classes are taken into account, the use of within-class ability grouping and instructional tailoring has no effect on average eighth-grade mathematics achievement; that is, average mathematics achievement in classes in which these practices are used extensively is not significantly different from that in classes in which these practices are not used at all. Since grouping and tailoring have no effect on average mathematics achievement across classes, what does? The factors affecting average achievement are class ability and opportunity-to-learn; that is, the ability level of the class as a whole and the amount of exposure to the curriculum. Classes that have high average achievement are those in which previous achievement and the exposure to content are high; these factors, and not whether grouping and tailoring are used, explain differences in average achievement across classes. Other contexts--the proportion of minorities and males within the class, the average SES and motivation levels, and the range of ability within the class--have no impact. Thus, in terms of excellence, neither ability grouping nor instructional tailoring would be considered more effective than their nonuse.

While it was not possible to determine directly if the effects of within-class ability grouping and instructional tailoring had a differential effect on students assigned to different levels of ability groups within a classroom, it was possible to determine whether either or both of these practices had an effect on the link between previous and subsequent achievement; in other words, whether ability grouping and instructional tailoring have an effect on the gap in subsequent achievement between high and low achievers? Classes vary in the strength of this relationship and this study examined the extent to which grouping and tailoring explained the variation in this relationship.

From the results of this study, one can conclude that the use of within-class ability grouping has no effect on the link between previous and subsequent eighth-grade mathematics achievement and the extent to which it varies across classes. That is, the gap between high and low achievers in posttest mathematics achievement was essentially the same regardless of the extent to which ability grouping was used in the classroom. Although classes vary in the strength of this relationship, the use of ability grouping does not explain this variation. However, one can conclude from the results that, over and above the effects of other characteristics of the class, instructional tailoring does have an effect on the link between previous and subsequent achievement. In classes in which tailoring was used more extensively, the relationship was stronger and the posttest mathematics achievement gap between high and low achievers was greater than it was in classes in which tailoring was not used. That is, classes vary in the strength of the relationship between previous and subsequent achievement and the extent to which tailoring was used explain some of this variation.

In addition to instructional tailoring, what factors affect the relationship between previous and subsequent achievement? Class ability and opportunity-to-learn levels have an effect on this relationship while the other contextual variables do not. That is, classes that have greater gaps in posttest mathematics achievement between high and low achievers are those in which previous achievement levels are high, exposure to content is high, and tailoring is used extensively. The extent

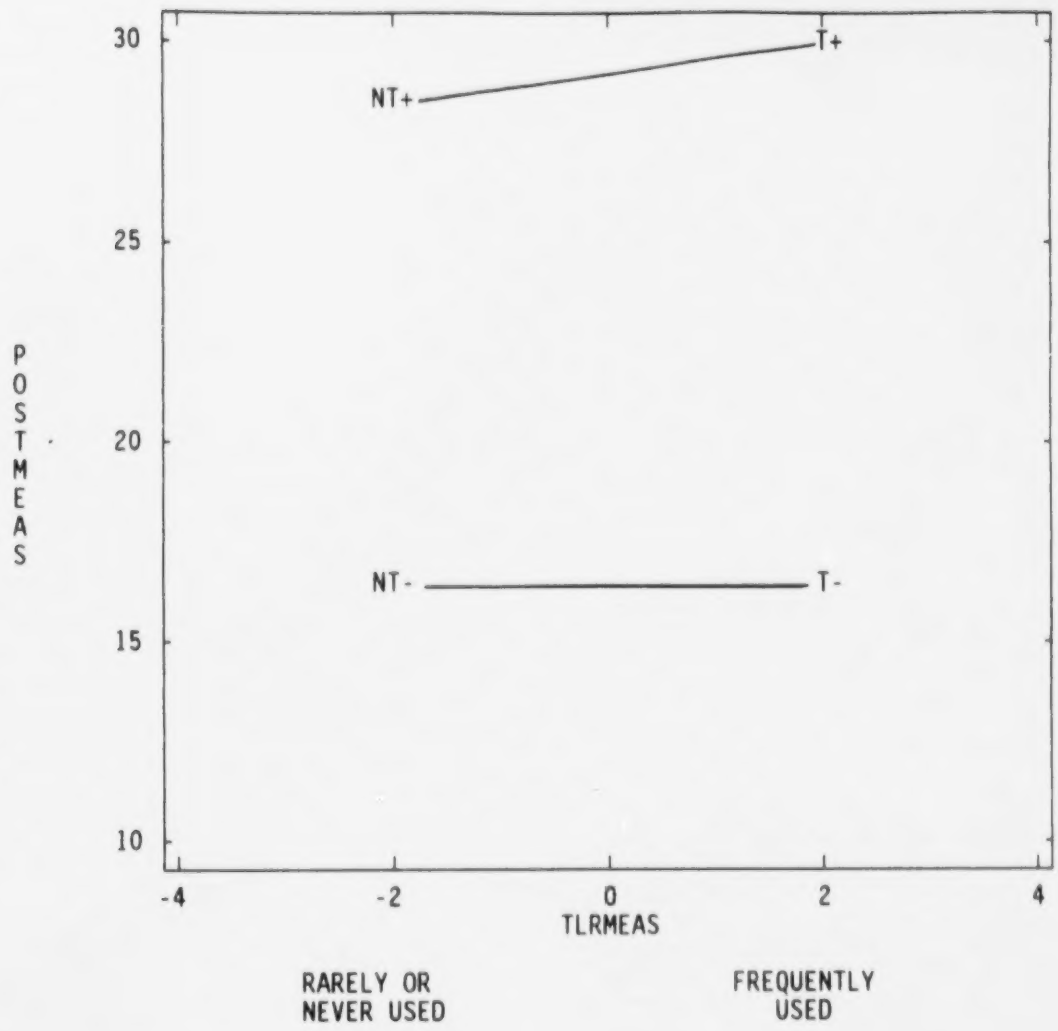


Figure 6. Posttest and Adjusted Posttest Measures for Different Levels of Tailoring for High and Low Achievers

to which tailoring (but not grouping) is used within the class explains some of the variation in the relationship between previous and subsequent achievement across classes. When equity is viewed as being equally effective for both high and low achievers, both ability grouping and instructional tailoring would be considered equitable.

What conclusions regarding the effectiveness and equitability of ability grouping and instructional tailoring can be drawn from this study? It appears that both within-class ability grouping and instructional tailoring, while not more effective, can be considered as effective as their nonuse. Ability grouping and instructional tailoring are as effective as their nonuse because students in eighth-grade mathematics classes in which they are used, and where the ability level and opportunity-to-learn level of the class are taken into account, have essentially the same level of mathematics achievement as those in classes where they are not used.

When equity is viewed in terms of being equally effective for high and low achievers, the conclusions are similar. If high achievers in both types of classes (those in which tailoring is used extensively and those in which tailoring is not used) were comparable in achievement at the beginning of the year, high achievers in classes in which tailoring was used extensively would perform better than those in classes in which tailoring was not used. If low achievers in both types of classes were comparable in pretest achievement, low achievers in classes in which tailoring was used extensively would perform as well as those in classes in which tailoring was not used. That is, tailoring benefits high achievers without adversely impacting the achievement of low achievers. Therefore, tailoring would be considered equitable.

Although the effects of instructional tailoring have not been previously studied separately from the effects of ability grouping, it's clear from this study that they need to be. Decisions made regarding the elimination of ability grouping would be made on incomplete information if the effects of instructional tailoring were not taken into account. While, according to this study, the use of ability grouping alone can be considered as effective and equitable, it is usually not used alone but in conjunction with instructional tailoring. Therefore, the equitability of instructional tailoring needs to be taken into account in deciding whether or not to discontinue the use of ability grouping. If the goal is to maximize student performance, regardless of whether students are high or low achievers, tailoring would be considered as equitable as nontailoring in that both groups perform as well if not better when tailoring is used. Since the goal of maximizing performance for all students could be considered the typical definition of equity, the conclusions regarding the effect of instructional tailoring are made on these results. It should be kept in mind that these results are based on the use of within-class ability grouping in eighth-grade mathematics classes and its effects, and the effects of instructional tailoring, may be different at different grade levels or in different subject areas.

Taking all of these issues into account, if recommendations were to be made regarding whether or not to use within-class ability grouping and instructional tailoring in eighth-grade mathematics classes, these recommendations would be as follows:

1) Instructional tailoring should be used in mathematics instruction; it allows high achievers to perform to their maximum while not having a negative effect on low achievers. The extent to which one groups students by ability doesn't appear to make a difference although ability grouping is typically the mechanism by which teachers tailor instruction.

2) Within-class ability grouping should be used instead of between-class grouping for mathematics instruction. While neither have an effect on average achievement, it appears that the effect of between-class grouping is negative for low achievers which is not the case when within-class grouping is used. Further, within-class grouping does not socially or academically isolate students to the same extent as does between-class grouping.

References

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon Jr, R.T.: *An Introduction to Hierarchical Linear Modeling: Computer Program and Users' Guide* (Version 2.0). Chicago, University of Chicago Press, 1988.

Eder, D.: Ability grouping as a self-fulfilling prophecy: A micro-analysis of teacher-student interaction. *Sociology of Education* 54:151-162, 1981.

Gamoran, A.: Organization, instruction, and the effects of ability grouping: Comment on Slavin's "best-evidence synthesis." *Review of Educational Research* 57(3):341-345, 1987.

Gamoran, A.: Schooling and achievement: Additive versus interactive models. In: *Schools, Classrooms, and Pupils: International Studies of Schooling From a Multilevel Perspective*, eds. S.W. Raudenbush and J.D. Willms, pp. 37-51. Orlando FL, The World Bank, 1991.

Good, T.L., and Marshall, S.: Do students learn more in heterogeneous or homogeneous groups? In: *The Social Context of Instruction: Group Organization and Group Process*, eds. P.P. Peterson, L.C. Wilkinson, and M.T. Hallinan, pp. 15-38. New York, Academic Press, 1984.

Hallinan, M.T., and Sorensen, A.B.: Ability grouping and sex differences in mathematics achievement. *Sociology of Education* 60:63-72, 1987.

Oakes, J.: *Keeping Track: How Schools Structure Inequality*. New Haven CT: Yale University Press, 1985.

Oakes, J., Gamoran, A. and Page, R.: Curriculum differentiation: Opportunities outcomes, and meanings. In: *Handbook of Research on Curriculum*, ed. P.W. Jackson, pp. 570-608. New York: Macmillan, 1992.

Page, R.: *Ambiguous Customs: An Ethnography of Lower-track Classrooms*. New York, Teachers College Press, 1992.

Raudenbush, S.W., and Bryk, A.S.: A hierarchical model for studying school effects. *Sociology of Education* 59:1-17, 1986.

Rosenbaum, J.E.: Social implications of educational grouping. In: *Review of Research in Education, Volume 8*, ed. D.C. Berliner, pp. 361-401. Washington DC, The American Educational Research Association, 1980.

Slavin, R.E.: Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research* 57(3):293-336, 1987.

Sorensen, A.A., and Hallinan, M.T.: A reconceptualization of school effects. *Sociology of Education* 50:273-289, 1977.

Sorensen, A.B., and Hallinan, M.T.: Effects of ability grouping on growth in academic achievement. *American Educational Research Journal* 23(4):519-542, 1986.

Westbury, I., Caroll, C., and Thalathoti, V.: *Second International Mathematics Study: United States Population A - Technical Summary*. International Association for the Evaluation of Educational Achievement. Urbana IL, University of Illinois at Urbana/Champaign College of Education, 1989.